
Article:
Jager, Tjalling and Ashauer, Roman orcid.org/0000-0002-9579-8793 (2018) How to evaluate the quality of toxicokinetic-Toxicodynamic models in the context of environmental risk assessment. Integrated Environmental Assessment and Management. pp. 604-614. ISSN 1551-3793

https://doi.org/10.1002/ieam.2026

How to Evaluate the Quality of Toxicokinetic-Toxicodynamic Models in the Context of Environmental Risk Assessment

Running head: How to evaluate the quality of TKTD models

Tjalling Jager,*† and Roman Ashauer‡

† DEBtox Research, NL-3731 DN, De Bilt, The Netherlands. tjalling@debtox.nl
‡ Environment Department, University of York, Heslington, York YO10 5NG, United Kingdom. roman.ashauer@york.ac.uk
* To whom correspondence may be addressed.

Editor's Note: "This is 1 of 10 articles generated from the session "Predictive models in ecotoxicology: bridging the gap between scientific progress and regulatory applicability," presented at the 27th SETAC Europe Annual Meeting (May 2017, Brussels, Belgium). The session considered approaches used in ecotoxicology for understanding and predicting the effects of chemicals, from QSAR to ecological modelling. This series aims to critically analyze and debate application examples and future developments to increase the acceptability of predictive models by regulators, managers, NGOs, and other stakeholders."

ABSTRACT

Environmental risk assessment (ERA) of chemicals relies on the combination of exposure and effects assessment. Exposure concentrations are commonly estimated using mechanistic fate models, but the effects side is restricted to descriptive statistical treatment of toxicity data. Mechanistic effect models are gaining interest in a regulatory context, which has also sparked discussions on model quality and good-modelling practice. Proposals for good-modelling practice of effect models currently focus very much on population and community models, whereas effects models also exist at the individual level, falling into the category of toxicokinetic-toxicodynamic (TKTD) models. In contrast to the higher-level models, TKTD models are usually completely parameterised by fitting them to experimental data. In fact, one of their explicit aims is to replace descriptive methods for data analysis. Furthermore, the construction of these models does not fit into an orderly modelling cycle, as most TKTD models have been under continuous development for decades, and are being applied by many different research groups, for many different purposes. These aspects have considerable consequences for the application of frameworks for model evaluation. For example, classical sensitivity analysis becomes rather meaningless when all model parameters are fitted to a data set. We illustrate these issues with the General Unified Threshold model for Survival (GUTS), relate them to the quality issues for currently-used models in ERA, and provide recommendations for the evaluation of TKTD models and their analyses.

Keywords: TKTD, GUTS, model quality, good-modelling practice, risk assessment

INTRODUCTION

Environmental risk assessment (ERA) is concerned with assessing the environmental impacts of human activitities, such as the release of chemicals. In general, chemical risk assessment comprises an exposure assessment and an effects assessment, but whereas the former relies heavily on application of mechanistic models, the latter is based on descriptive procedures such as hypothesis testing and curve fitting on the results of standard toxicity tests. Such descriptive methods have a range of problems associated with them (Laskowski 1995; Jager 2011). Most importantly: they provide biassed estimates by focussing only on effects after a standardised exposure time, and do not allow for useful extrapolations to other exposure scenarios. Mechanistic effect models do exist, and are rapidly gaining interest in the context of ERA for chemicals (Grimm et al. 2009; Preuss et al. 2009b; Hommen et al. 2010; Hommen et al. 2016), specifically for plant-protection products (PPPs). As the European Food Safety Authority (EFSA 2013) puts it: "It is expected that mechanistic effect models at all levels of biological organisation will be used to support the RA of PPPs in the future." This interest in effect models has also sparked a discussion regarding the quality of these models, and how to document and evaluate it.

A general guidance for the evaluation of models to be used for ERA was produced in 2009 by the US Environmental Protection Agency (US-EPA 2009). Specifically for effects models, two frameworks were presented more recently: the transparent and comprehensive ecological modelling (TRACE) documentation (Schmolke et al. 2010; Grimm et al. 2014), and EFSA's scientific opinion on good-modelling practice for effect models in the context of PPPs (EFSA 2014). Whilst the recommendations given in these frameworks are generally sensible, it is not so clear whether they can be applied in the same way to models at different levels of biological organisation. The effects models at the individual level fall into the category of toxicokinetic-toxicodynamic (TKTD) models. These models differ in key aspects from ecological population or system models. Specifically, TKTD models are usually completely calibrated with case-specific data, and they have a long history of conceptual development and application by many different research groups. What are the consequences of these differences for quality evaluation of TKTD models? In this paper, we will explore this question and its repercussions in more detail, using the General Unified Threshold model for Survival (GUTS, Jager et al. 2011) as an illustration. GUTS is a typical example for TKTD modelling in ecotoxicology, and the issues we discuss for this model hold for other TKTD models as well. Furthermore, we will compare these issues for quality evaluation of TKTD models to those for models that are currently routinely used in ERA, and provide a way forward.

It is good to stress up front that model quality cannot be assessed in isolation as it is tightly linked to the intended use of the model. As the US Environmental Protection Agency puts it (US-EPA 2009): "Quality is an attribute of models that is meaningful only within the context of a specific model application. Determining whether a model serves its intended purpose involves in-depth discussions between model developers and the users responsible for applying the model to a particular problem." Model quality evaluation is thus best viewed as an analysis of whether a model (or an application of a specific model) is fit-for-purpose. Therefore, we start with a short general introduction into TKTD modelling and highlight the purposes for which these models can be applied.

TKTD GENERAL INTRODUCTION

Mechanistic effect models at the individual level belong to the category of TKTD models. These models combine a toxicokinetic (TK) and a toxicodynamic (TD) module to provide a link between external concentrations and the effects on relevant endpoints over

time. In the context of ERA, relevant individual-level endpoints are those with a direct link to population performance, i.e., survival, growth and reproduction. In contrast to risk assessment for human health, the focus lies on the protection of populations and ecosystems, and not the well-being of individuals per se (see Van Leeuwen and Vermeire 2007). Both the TK and TD module explicitly consider the factor 'time', in a mechanistic manner. Therefore, TKTD models allow the external concentration and other environmental conditions to vary over time, and the resulting effects on the endpoints are predicted as they change over time (for some models even over the full life cycle of the organism).

Most TKTD models have been developed with the explicit aim to be generic: the same model can be used for many species and chemicals. Application of these models starts with calibrating the model to toxicity data for the chemical and species of interest. The difference between different species and chemicals is thus reflected in different values of the model parameters, and not (or much less) in model structure. Coupled to the fact that TKTD models are based on a representation of the underlying mechanisms, this allows for useful comparisons between species and chemicals, and for extracting general (and hopefully predictive) patterns. These models are usually developed with standard ecotoxicity data in mind, which implies that they can be calibrated using observations on individual-level endpoints such as survival and reproduction, without requiring measurements at the sub-individual level (e.g., histology or body residues).

Owing to these preconditions, TKTD models are relatively simple, and based on rather abstract representations of the underlying mechanisms. For example, GUTS models (Jager et al. 2011) view mortality as a stochastic process (lumping an enormous number of processes into a chance process), and Dynamic Energy Budget models (DEB, see Sousa et al. 2008) focus on few idealised biomass components (e.g., structure and reserve) and lumped energy fluxes such as total maintenance costs. Such a huge simplification of biology and toxicology implies that many potentially relevant processes are ignored, but is needed to accommodate limited data sets and to keep the models generic and transparent. More general discussion on TKTD modelling and its advantages can be found elsewhere (Jager et al. 2006; Ashauer and Escher 2010).

TKTD models have a substantial history in science, where they are used to understand and explain the effects of toxicants (and other stresses) on the life-history traits of a species. Although TKTD models are currently not routinely used in ERA, their potential for application has been recognised in several frameworks. For example, they have been included in ISO/OECD guidance (OECD 2006) as biology-based methods, and EFSA incorporated them in their risk assessment scheme to complement laboratory studies with single species in tier 2 (EFSA 2013). As stated in the introduction, model quality is tightly bound to the intended use of the model. TKTD models may be used for various tasks in ERA, which we list in Table 1. In this list, TKTD models are compared to traditional dose-response curves as they can serve the same purposes (but then with greater power and flexibility), and can fulfil additional purposes that traditional models cannot.

The TKTD model that we will use as a typical example is the General Unified Threshold model for Survival (GUTS, Jager et al. 2011). GUTS is not a single model, but rather a framework from which specific models can be derived in a consistent manner (by fixing parameters to certain values). It is limited to effects on mortality or immobility, and builds on a long history of survival analysis in science. In fact, almost all survival models that have been proposed in ecotoxicology can now be viewed as special cases. GUTS is a collaborative result, evolving from efforts to harmonise research in the field of survival modelling, in order to make more efficient progress, both in science and in applied contexts such as ERA. This is in effect comparable to the development of a consensus model for multi-media chemical fate

to calculate long-range transport potential, which is commended in the US-EPA guidance (US-EPA 2009, p.23, box 6).

To illustrate what GUTS does, we provide an example in Figure 1 for one special case of the model (the reduced stochastic death model) for propiconazole in the amphipod *Gammarus pulex* (data from Nyman et al. 2012). This is a typical example for the first use of TKTD models, as listed in Table 1: analysing rather standard toxicity data for survival over time, using all data in a single analysis. GUTS requires just four parameters for the entire data set, and all of these parameters are calibrated: all parameters obtain their value by fitting the model to data at the same level as the model's output (i.e., model predictions for survival probability are compared to observations on survival in the experimental cohort).

WHAT MAKES TKTD MODELS DIFFERENT

In applying good-modelling frameworks to TKTD models, a number of problems arise in practice. Firstly, these frameworks implicitly focus on a specific category of effect models, namely the models at the population level and higher (community, ecosystem, landscape). The criteria they propose cannot be used in the same way for effects models at the individual level, such as TKTD models. Specifically, these frameworks request reconstruction and documentation of every step in the modelling cycle (Fig. 2), which is impossible for generic model concepts with a long history and broad distribution in the scientific community. Furthermore, TKTD models are, unlike population models, always fitted to data. These issues will be discussed in more detail below, followed by their consequences for mainstays of good-modelling frameworks: sensitivity/uncertainty analysis and validation. It is good to stress already here that these problems are not unique for TKTD models; they are shared by the effect models that are currently used in ERA (discussed in more detail later).

The Modelling Cycle

The TRACE framework (Schmolke et al. 2010; Grimm et al. 2014) and EFSA opinion (EFSA 2014) emphasise the modelling cycle (simplified version shown in Fig. 2), and documentation of that cycle: "Model development should follow the modelling cycle, in which every step has to be fully documented …" (EFSA 2014). Such a cycle may be a realistic representation when a model is built, largely from scratch, by a single research group (or even a single person), to address a specific problem. This situation is perhaps representative for many models at higher levels of biological organisation, although even these models are not built entirely from scratch; most will rest on well-tested foundations such as matrix and individual-based modelling. However, on this foundation, a specific model structure is developed and parameterised for a specific species (sometimes also a specific chemical), a specific environmental setting, and for a specific purpose. This particular model structure, with its general parameterisation, is then usually referred to as 'the model', and can be documented in the form of a modelling cycle as in Figure 2. TKTD models, in contrast, are truly generic: one model structure is applied to many species, chemicals, environmental settings and purposes. The most prominent model concepts have a long and convoluted history in ecotoxicology. One model is typically applied, tested and further developed by many different research groups, working largely independently, but regularly borrowing and modifying elements of each other's work. The same model structure will have been implemented in different software applications, used for different species-chemical combinations, and applied to address very different questions. As a consequence, there is no modelling cycle for a particular model; instead, we are looking at a complex modelling web.

Taking the example of GUTS, the model structure has not been developed from scratch in response to a specific problem formulated in an ERA context; it is based on a century of work on survival modelling, which had led to a wide range of (seemingly unconnected) models. GUTS places all these developments into a single framework, and thus builds on the insights gained with all of these models. Specific cases of this framework have been implemented in different programming languages (Matlab, R, Delphi, etc.), by different groups, independently. When we have selected a particular special case of GUTS, and a specific software implementation, it can then be applied to a wide range of data sets for different species and chemicals, and to a variety of problems. There is also no closing of the modelling cycle as in Figure 2: a single application will not generally lead to adaptation of the conceptual GUTS model, although everyone is free to develop their own version of GUTS, as there is no central coordination.

Clearly, with GUTS, we do not see a single model going through a cycle. We have (at least) three different stages of model development, with an ever expanding number of 'models': there is one conceptual GUTS framework (with a limited number of special cases, stage 1), a wider range of software implementations (probably a few dozen, stage 2), and a huge number of applications to specific data sets and specific questions (possibly thousands, stage 3). This expansion of models is illustrated in Figure 3. GUTS builds on work that has been done on other models for survival, and includes the one-compartment TK model with first-order kinetics as a module, which has its own (even more distant) historical roots. This can be considered stage 0 of the model.

The modelling cycle is an overly-idealised picture, and a picture that actually fails in representing the most interesting models for ERA: generic models that have a long history and widespread support in the scientific community. Even though there is no modelling cycle for TKTD models such as GUTS, the elements of the cycle in Figure 2 still make sense. However, they apply to different stages of the modelling (Fig. 3). For example, there is no point in asking for an uncertainty analysis of GUTS in general (explained in detail in the section 'Sensitivity And Uncertainty Analysis' below); such an analysis would only make sense after a specific case of GUTS has been calibrated to a data set, and thus for an application (stage 3). Furthermore, it would be superfluous to document the quality of the GUTS concepts for each application again and again. It is more useful to separate the evaluation of the conceptual model (stage 1) from the evaluation of the software (stage 2), from the evaluation of each specific application of the model (stage 3, as in Fig. 3). Specific questions relate to each stage, and require a different type of evaluation and documentation (which should refer to, rather than repeat, the evaluation/documentation of the preceding stages).

Difference In Parameterisation/Calibration

In the previous section, we discussed the fact that there is no modelling cycle in sight for most TKTD models, as might be more easily discerned for population/community models. Another (though related) difference with such higher-level models is that TKTD models are almost always directly fitted to data; data of the same type as the model output. GUTS is a model for the survival probability of individuals over time, and is fitted to observations on survival for a test cohort over time. And not just some of the parameters are fitted, but generally all of them. In fact, one of the main applications of TKTD models is as a tool for data analysis (Table 1). This implies that a number of traditional evaluation criteria, such as sensitivity/uncertainty analysis and validation, need to be reconsidered. This is most clearly illustrated by contrasting the situation for TKTD models to that for population models.

Population models are rarely fitted to measured population densities; exceptions are the simple exponential and logistic growth models that are used for algae and duckweed (and for

these models, the same problems as for TKTD models apply). The number of model parameters is usually way too large to be determined solely from population data, and, more fundamentally, if such extensive data sets existed, there would be little need for the model in the first place (modelling is used to predict population impacts in the absence of relevant observations). Instead, model parameters in a population model obtain their value from a parameterisation process that does not involve the model itself (Fig. 4). For example, a population model might require a feeding rate for individuals, and we can look at experimental data on the feeding process to derive a relationship between ingestion rate and body length (e.g., Preuss et al. 2009a). After all parameters have received a value, we can run the model and obtain model predictions at the population level. Classical uncertainty/sensitivity analysis implies that we change the parameter values and investigate the response of the model prediction to those changes. We can change model parameters in a rather arbitrary manner (sensitivity analysis) or apply realistic distributions for them (uncertainty analysis). We can compare model predictions and independent observations (corroboration), and/or modify some parameter values to obtain a closer correspondence to observed patterns (calibration). Since most of the model parameters will be fixed for a range of applications (e.g., the properties of the species and the environment), they (and the data sets they are based on) are often treated as an integral part of the model. This is likely the underlying rationale for including data-dependent modelling steps (parameterisation, calibration, corroboration, sensitivity/uncertainty analysis) into the development cycle for a model (Fig. 2).

The situation sketched above for population models is very similar to the case for chemical fate models, which are obviously not parameterised by fitting them to measured environmental concentrations. For TKTD models, however, parameterisation is the same as calibration, which is the actual application of the model to fit a data set (Fig. 4). Observations of the same type as the model output are used to obtain the value for all of the model parameters, as was illustrated with the example of GUTS (Fig. 1). This is a fundamentally different way to use models, and the values of the model parameters (and the data sets used to derive them) cannot be considered as part of the model in any sense (unless we are willing to treat the result of each fitting exercise as a new model). For TKTD models, including the data-dependent modelling steps in an evaluation of the model concept therefore does not make any sense. On closer inspection, this also makes little sense for other models, such as those at the population level; it is best to explicitly separate the evaluation of the model concepts from that of its parameterisation. Therefore, the modelling cycle of Figure 2 needs to be replaced by a model expansion as in Figure 3. The fact that TKTD models are fully parameterised by calibration has important consequences for how we should look at several data-dependent components of good-modelling practice such as sensitivity/uncertainty analysis and validation.

Sensitivity And Uncertainty Analysis

Classical sensitivity and uncertainty analysis are hardly useful for TKTD models. To perform such analyses, the parameters first need to have received a value, or at least a reasonable range, and we need a scenario (e.g., exposure pattern and duration). Otherwise, we would need to change all parameters between zero and infinity, for all possible exposure patterns, which would be rather pointless. The parameters of TKTD models receive their value in a calibration to a particular data set. We thus have to start by fitting the model, as illustrated for GUTS in Figure 1, for the case of propiconazole in *G. pulex*. However, after fitting the model, what is the point in changing parameters to see how the model output changes? When we change the value of a parameter, we obviously change the goodness-of-fit, and that was already done by the optimisation routine. In fact, it is the fit on the data that informs us about the sensitivities and uncertainties in the model parameters: the (joint) confidence interval on

the parameters provides all of the relevant information. Parameters with a narrow confidence interval are sensitive, and not very uncertain (their value is clearly identified from the data set). Parameters with a wide confidence interval cannot be properly identified from this particular data set. They are either insensitive in the model (in this particular part of parameter space), and/or the data do not allow their identification.

Of course, after the model has been fitted, we can still perform a classical sensitivity and uncertainty analysis. As explained above, this does not yield much in terms of useful information, and furthermore, the sensitivity of each parameter will depend on the exposure concentration, the time of exposure, and even the exposure pattern (Ashauer et al. 2013), as well as on the specific type of model output selected. In population models, sensitivity analysis is useful for those parameters that are treated as part of the model (e.g., specifying individual behaviour or environmental conditions), and especially for those that are not fixed by relevant data. Such an analysis helps refining the model: we need to scrutinise the parameterisation for those parameters to which the model output is particularly sensitive, and can remove processes that are insensitive. However, in GUTS, all parameters are fitted simultaneously to a set of survival data (Fig. 1), and we cannot refine the parameterisation for an individual parameter as it is an integrated part of the model fit, and parameters tend to co-vary. If we don't like the parameter values or their confidence intervals, the only thing we can do is to perform a new toxicity experiment (which, with help of the model, can be designed to maximise the identification of certain parameters, if needed; Albert et al. 2012).

Instead of classical sensitivity and uncertainty analysis, application of TKTD models benefits most from a robust and coherent statistical treatment in model fitting. The confidence intervals on the parameters tell us how well they can be identified from this particular data set, and the joint confidence set can be used to propagate uncertainty in the model parameters to uncertainty in model predictions (for examples, see Jager and Zimmer 2012; Ashauer et al. 2016). It is important to stress that the approaches for model calibration and dealing with parameter uncertainty tell us nothing about the quality of the conceptual model. They are aspects of a specific model application (stage 3, see Fig. 3) in combination with specific calibration data.

Model Validation

Since the starting point of virtually all analyses with TKTD models is a calibration to the available data, this also raises the question of how to judge the degree of correspondence between model and reality. This process is often referred to as 'validation', but as models are necessarily simplifications of nature they can never be validated in its strict sense (Oreskes et al. 1994). It is therefore better to use 'corroboration' instead, to refer to the comparison of predictions to observations that were not used for calibration (see US-EPA 2009). Corroboration is an important element of modelling as it clarifies (and in the best case, quantifies) how well the model can represent reality. There are many limitations and pitfalls, for example, a model might perform well for the wrong reasons (Oreskes et al. 1994). Nevertheless, a range of corroboration studies can go a long way in embodying trust in the usefulness of the model concept (for a specific purpose).

The requirements for corroboration should be closely linked to the intended use of the model. If the TKTD model is to be used to derive a no-effect threshold or an ECx,t, the requirements will differ from the situation where the model is to be used to extrapolate to untested exposure scenarios or to different species (Table 1). In any case, verification is possible and very important: checking the consistency and realism of the underlying assumptions, and the translation into mathematics and computer code. However, corroborating the model's output with independent observations becomes rather awkward for the application

of TKTD models in data analysis. In a population model, the type of data that the model predicts (e.g., population abundance and structure over time) is not the type of data that is used to parameterise the model (see Fig. 4). Therefore, we can parameterise the model first, and compare its predictions to independent observations from a population experiment (for an example, see Preuss et al. 2009a). However, for TKTD models, parameterisation involves fitting the model to data for the endpoint that is being predicted. Taking the GUTS example in Figure 1, all of the experimental data are used for the parameterisation/calibration. We could repeat the experiment and compare the calibrated model to the new data set, but that would say more about the reproducibility of the experimental test than on the realism of the model. The only criterion we can use for 'validity' of the model in these applications is the goodness-of-fit of the model to the data. If the model fits well, and if it generally fits well on this type of data, that provides support for the model (but not a corroboration in the strict sense). However, goodness-of-fit has to be viewed in relation to the flexibility of the model to produce different patterns. A twenty-parameter TKTD model might provide a good fit to any data set you throw at it, but the goodness-of-fit will not provide support for the realism of the model concept anymore.

We do have the opportunity to corroborate TKTD models using independent data when we use the model to make extrapolations beyond the calibration data set. For example, we can use the calibrated GUTS model in Figure 1 to make predictions for effects at untested time points, untested exposure concentrations or exposure scenarios (e.g., pulse exposure), or untested environmental conditions. Subsequently, we can set up additional experiments to test those predictions. These are very useful exercises to clarify the accuracy and precision of the model predictions, but they have been quite rare so far for TKTD models (for GUTS, see Nyman et al. 2012; Ashauer et al. 2016). However, there are several stumbling blocks that must be considered. Firstly, performing corroboration studies is hampered by the way science is financed; after all, such exercises are not considered sufficiently novel for most funding agencies. Furthermore, corroboration will be most convincing if it covers the intended uses, and at this moment, it is unclear how these models are to be used in ERA (see the options in Table 1). And, finally, a corroboration study requires the model to be parameterised for a specific case (chemical, species, conditions, etc.). This implies that a lack of correspondence might be either caused by a failure of the model concept, or a failure in the parameterisation, or both. Like sensitivity/uncertainty analysis, corroboration is a data-dependent analysis, and thus part of an application (stage 3 in Fig. 3).

CURRENT MODELS FOR EFFECTS

Same Problems With Current Models

The points raised above are not unique for GUTS or for TKTD models in general; the same problems apply to classical dose-response analysis, TK models, and the simple exponential and logistic models for algal population or plant growth. These methods are also models, and have been routinely used in ERA for decades, so it makes sense to scrutinise them from a good-modelling perspective. Furthermore, looking at these more familiar models helps clarify the points made above for TKTD models.

Clearly, it is impossible to discern a modelling cycle (Fig. 2) for these models, and in many cases, it will even be difficult to trace their origins. Furthermore, few would attempt sensitivity and uncertainty analysis on these models; after fitting a dose-response curve to a data set, there is no point in varying the parameters (e.g., EC50 and slope parameter) to see how the curve changes. Once we have established an EC50 with a confidence interval, there is simply no need for additional analyses. The question of 'validity' also becomes rather trivial: how can we

establish the realism of the fitted dose-response curve or the EC50 that follows from it? Regarding evaluation of the model concepts, the dose-response curves have no underlying logic, and as OECD guidance (OECD 2006) puts it: "A statistical regression model itself does not have any meaning, and the choice of the model (expression) is largely arbitrary." For corroboration, we could redo the experiment, resulting in independent data, and see if the curve established earlier also goes through this data set. However, if the new data are very different from the calibrated model, it will also be very different from the first data set. Hence, the issue would be with the reproducibility of the test, and not the quality of the model or its calibration.

The situation is equivalent for the exponential growth model that is used to analyse toxicity data for algae and duckweed growth, and the one-compartment TK model that is used in analysing bioconcentration data. There is no modelling cycle, nobody asks for sensitivity or uncertainty analyses, it may be questioned whether these models (or their implementations) have been verified, and certainly they are never corroborated using independent data. As with TKTD models, the only criterion we can use for 'validity' is the goodness-of-fit of the model to a data set, and to this type of data in general. For these classical models, this is acceptable as they have a very limited range of curve shapes that they can produce; they will not fit just any data set. In fact, the log-logistic dose-response curve, exponential growth curve, and one-compartment TK model have only one curve shape, and the model parameters scale or shift this shape in the x- and y-direction.

Which Criteria Were Used To Select Current Models?

As explained in the previous section, models are already being used in the effects assessment of ERA, albeit rather crude ones. What kind of criteria were used to select these models for ERA in the first place? Certainly not the criteria put forward in good-modelling frameworks, as the currently-used models would not have passed. For these models, there is no modelling cycle, and therefore also no documentation of every step in their development and application. Criteria like sensitivity/uncertainty analysis and 'validation' have clearly never been an issue. Interestingly, there seem to be no criteria on the software implementations to be used, and even more striking: the test guidelines currently don't even specify which model to use, and which methods and software to fit them with. The guidance for acute fish toxicity testing (OECD 1992) states: "Normal statistical procedures are then employed to calculate the LC50 for the appropriate exposure period." The same is true for the Daphnia reproduction test (OECD 2012): "ECx-values, including their associated lower and upper confidence limits, are calculated using appropriate statistical methods (e.g. logistic or Weibull function, trimmed Spearman-Karber method, or simple interpolation)."

The most important criterion used in the past to select these models for ERA was probably the fact that there were no alternatives at the time, and that these approaches had a broad acceptance in ecotoxicology and among the stakeholders. This is understandable as the effectivity of ERA rests on acceptability in science, industry and society. Furthermore, models that have managed to gain such broad acceptance have probably done so on the basis of their (perceived) appropriateness for the problem at hand. This illustrates that support for a model may be more important for its use in ERA than whether it passes all of the criteria for good modelling. However, it also shows a clear mismatch between the criteria proposed for new effects models and the demands that are placed on currently used methods.

WAY FORWARD

Modifying Criteria For TKTD Models

TKTD models share a number of features with the models currently used in effects assessment by which they fundamentally differ from other types of models such as population/community models and fate models. Most strikingly: it is impossible to identify a modelling cycle (as in Fig. 2) for TKTD models, and they are parameterised directly and completely by fitting them to data (Fig. 4). As a consequence, there is a mismatch between TKTD models and the frameworks for good-modelling practice that are currently being put forward. To remedy this mismatch, the model evaluation needs to be broken down in (at least) three separate stages: the conceptual model, the model implementation, and the model application. One conceptual TKTD model (such as GUTS) will generally have many implementations, and even more applications (Fig. 3). Each stage has a set of evaluation points that is relevant for that stage only. In Table 2, we adopted and modified typical items for a good-modelling evaluation, and categorised them into the different stages that 'the model' encompasses. Quality issues only propagate downstream: an issue with the conceptual model will influence all applications that follow from it, but an issue with an application does not affect the quality of the conceptual model.

Especially within this last stage, some criteria will need to be modified to suit TKTD models. For example, classical sensitivity/uncertainty analysis is rather meaningless, and is best replaced by a proper derivation of confidence intervals, and propagation of parameter uncertainties from the fit to model predictions. Output corroboration is also part of the application as one can only compare a model to independent observations after it has been parameterised for a specific species and chemical. For TKTD models, corroboration needs to be reconsidered as these models are fitted to data (fully comparable to current use of dose-response curves). One element of 'validation' is goodness-of-fit and how well the model generally captures patterns in the kind of data that it is used for. However, goodness-of-fit has to be viewed in relation to the flexibility of the model to produce different patterns. When a TKTD model is used to make predictions beyond the range of the calibration data, proper output corroboration is possible (i.e., comparing model predictions to independent data, not used for calibration). At this moment, such studies have been rare for TKTD models, and they would need to closely match the intended use of the model in ERA to be meaningful.

Such a sub-division in stages is not only useful for models at the individual level, but also for other models to be used in ERA, especially those with a wide distribution in science. For most population models, it would be easy to evaluate stage 1 and 2 together as it seems that most models only have one software implementation. Models are not particularly useful for ERA if they only have a single application, so stage 3 would have to be treated separately just as for the TKTD models. For population models, however, not all of the model parameters are case specific; most will probably be kept constant for a range of application cases, such as the environmental setting or the behaviour of the focal species. Therefore, the values of these parameters are often treated as part of 'the model'. In our opinion, it is more useful to strictly separate the model from its parameterisation, the latter confined stage 3. Stage 3 would then be the only stage involving the interplay between data, parameter values, and model results. In the ERA context, however, it would make sense to separate stage 3 into a case-specific part (chemical and application specific) and a more general part. Confining all data-driven activities to stage 3 would clarify that any validation exercise or sensitivity/uncertainty analysis always relates to a particular parameterisation of the model.

How To Use Criteria For Model Quality

In our opinion, evaluation questions for model quality (such as those in Table 2) should not be used as absolute pass/fail criteria. Models (or their implementations/applications) that perform poorly on some of these criteria might still provide useful information for ERA, as the

current methods for effects assessment testify. What is the purpose of a model quality evaluation? Firstly, it is important to establish how each model scores on a set of quality criteria, and to clarify what their strengths and limitations are. This knowledge will help risk assessors to weigh the evidence presented by the model with other lines of evidence, and select appropriate assessment factors. Secondly, such scores can help select the most appropriate models for ERA.

The decision to use a new method (such as a model) in ERA should always be a comparison between the new method and the current (default) method. Proposing stringent criteria for new methods, without scrutinising the existing method (that the new method is supposed to replace) in the same manner, is bound to yield biased decisions. In selecting the most appropriate model for ERA, it is therefore essential to compare all contenders on an equal footing. This implies that the models and methods that are currently routinely used in ERA (such as dose-response curves) should also be judged by same criteria as put forward for newcomers such as TKTD models. If all items in Table 2 need to be evaluated and documented for each (application of a) TKTD model, why forego on such actions for dose-response models? Furthermore, it would be good to harmonise the demands on effects models to those for fate models as well. As an example, for fate models, propagation of uncertainties is not standard practice, and certainly not for each application. If such analyses are deemed essential for effects models, they should also be requested for fate models.

Besides quality criteria as in Table 2, there might be important reasons to prefer the current methods over mechanistic alternatives (e.g., political or regulatory constraints). These criteria should, however, be formalised and explicitly incorporated into the good-modelling frameworks for regulatory purposes; there is no point in trying to improve models to meet a list of criteria when additional implicit criteria are being used in practice.

Considering Modularity In Models

TKTD models are by definition a combination of (at least) two models: a TK model and a TD model. The TK models used are also applied by themselves to analyse body-residue data over time. In turn, TKTD models are sometimes used as building blocks in population models (for GUTS, see e.g., Dohmen et al. 2016). Clearly, individual models can be used as modules in more complex models. This modularity raises a number of issues for model evaluation. Firstly, it is more efficient to evaluate modules than it is to evaluate models, because modules can be re-used in many different models. The EFSA opinion (EFSA 2014) also recognises this: “Different models could be also combined together to form linked systems; in this case the individual models need to be considered as separate entities before considering the method by which they are integrated to form the whole model being used for a specific case.” The second issue with modularity is perhaps a less obvious one: it makes little sense to use different modules for the same process in different parts of the risk assessment. For example, if the one-compartment TK model is an appropriate model for body residues over time, it makes sense to use it in TKTD models as well (GUTS applies the same TK model in a TKTD context). If GUTS is an appropriate model for survival of individuals over time, it makes sense to apply it for that purpose as a building block in population models as well, instead of reverting to a static dose-response curve for the higher-level models. Using the same modules throughout the ERA process is not only logical but will also improve efficiency and consistency.

Agreed Models Or Modules

In risk assessment of PPPs in Europe, the demand for formal evaluation of model quality seems to be mainly driven by the fact that applicants can submit a calculation with any model

of their choice into the dossier. The risk assessors then subsequently faces the task to assess whether this model calculation has sufficient quality or not, based on the information provided by the applicant. This is an untenable situation as risk assessors cannot be expected to be (or become) specialists on every model. Instead of diverting the burden of proof to the model developers (who might not have an interest in ERA at all), a far more efficient way forward is to arrive at a set of agreed effects models (or better: modules), as proposed by others as well (EFSA 2014; Hommen et al. 2016). For a small selection of models, it will be straightforward to produce extensive documentation, user-friendly software (including methods for sensitivity and uncertainty analysis, if required), and dedicated verification and corroboration studies to fill the most important gaps that inevitably exist for every model. Furthermore, it would be relatively easy for all stakeholders to build up sufficient expertise with these models to allow for critical evaluation of model applications in a dossier. And finally, such a selection of models can also provide a focus for further scientific work: currently, researchers have a tendency to develop new models (stimulated by funding agencies and journals pressing for novelty), rather than applying and testing existing models or modelling frameworks. The experience with GUTS has shown us that much more progress can be made by collaborating and pooling resources around a single framework.

Acknowledgement – This research was financially supported by CEFIC-LRI, project ECO39: Review, ring-test and guidance for TKTD modelling (http://cefic-lri.org/projects/eco39-review-ring-test-and-guidance-for-tktd-modelling/ and http://debtox.nl/projects/project_guts.html).

Data accessibility – The data used in Figure 1 are available as supplementary material of Nyman et al, doi:10.1007/s10646-012-0917-0.

REFERENCES

Albert C, Ashauer R, Künsch HR, Reichert P. 2012. Bayesian experimental design for a toxicokinetic-toxicodynamic model. *J Stat Plan Infer* 142:263-275.

Ashauer R, Albert C, Augustine S, Cedergreen N, Charles S, Ducrot V, Focks A, Gabsi F, Gergs A, Goussen B, Jager T, Kramer NI, Nyman AM, Poulsen V, Reichenberger S, Schäfer RB, Van den Brink PJ, Veltman K, Vogel S, Zimmer EI, Preuss TG. 2016. Modelling survival: exposure pattern, species sensitivity and uncertainty. *Scientific Reports* 6:11.

Ashauer R, Escher BI. 2010. Advantages of toxicokinetic and toxicodynamic modelling in aquatic ecotoxicology and risk assessment. *J Environ Monit* 12:2056-2061.

Ashauer R, Thorbek P, Warinton JS, Wheeler JR, Maund S. 2013. A method to predict and understand fish survival under dynamic chemical stress using standard ecotoxicity data. *Environ Toxicol Chem* 32:954-965.

Dohmen GP, Preuss TG, Hamer M, Galic N, Strauss T, Van den Brink PJ, De Laender F, Bopp S. 2016. Population-level effects and recovery of aquatic invertebrates after multiple applications of an insecticide. *Integr Environ Assess Manag* 12:67-81.

EFSA. 2013. Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA journal* 11:3290.

EFSA. 2014. Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA journal* 12:3589.

Grimm V, Ashauer R, Forbes V, Hommen U, Preuss TG, Schmidt A, van den Brink PJ, Wogram J, Thorbek P. 2009. CREAM: a European project on mechanistic effect

models for ecological risk assessment of chemicals. *Environ Sci Pollut Res* 16:614-617.

Grimm V, Augusiak J, Focks A, Frank BM, Gabsi F, Johnston ASA, Liu C, Martin BT, Meli M, Radchuk V, Thorbek P, Railsback SF. 2014. Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecol Modell* 280:129-139.

Hommen U, Baveco JM, Galic N, Van den Brink PJ. 2010. Potential application of ecological models in the European environmental risk assessment of chemicals: review of protection goals in EU directives and regulations. *Integr Environ Assess Manag* 6:325-337.

Hommen U, Forbes V, Grimm V, Preuss TG, Thorbek P, Ducrot V. 2016. How to use mechanistic effect models in environmental risk assessment of pesticides: case studies and recommendations from the SETAC workshop MODELINK. *Integr Environ Assess Manag* 12:21-31.

Jager T. 2011. Some good reasons to ban EC*x* and related concepts in ecotoxicology. *Environ Sci Technol* 45:8180-8181.

Jager T, Albert C, Preuss TG, Ashauer R. 2011. General Unified Threshold model of Survival - a toxicokinetic-toxicodynamic framework for ecotoxicology. *Environ Sci Technol* 45:2529-2540.

Jager T, Heugens EHW, Kooijman SALM. 2006. Making sense of ecotoxicological test results: towards application of process-based models. *Ecotoxicology* 15:305-314.

Jager T, Zimmer EI. 2012. Simplified Dynamic Energy Budget model for analysing ecotoxicity data. *Ecol Modell* 225:74-81.

Laskowski R. 1995. Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *Oikos* 73:140-144.

Nyman AM, Schirmer K, Ashauer R. 2012. Toxicokinetic-toxicodynamic modelling of survival of *Gammarus pulex* in multiple pulse exposures to propiconazole: model assumptions, calibration data requirements and predictive power. *Ecotoxicology* 21:1828-1840.

OECD. 1992. OECD guideline for testing of chemicals 203. Fish, acute toxicity test. Organisation for Economic Cooperation and Development (OECD), Paris, France

OECD. 2006. Current approaches in the statistical analysis of ecotoxicity data: a guidance to application, Series on Testing and Assessment, No. 54. Organisation for Economic Cooperation and Development (OECD), Paris, France

OECD. 2012. OECD guideline for testing of chemicals 211. *Daphnia magna* reproduction test. Organisation for Economic Cooperation and Development (OECD), Paris, France

Oreskes N, Shrader-Frechette K, Belitz K. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263:641-646.

Preuss TG, Hammers-Wirtz M, Hommen U, Rubach MN, Ratte HT. 2009a. Development and validatlion of an individual based *Daphnia magna* population model: The influence of crowding on population dynamics. *Ecol Modell* 220:310-329.

Preuss TG, Hommen U, Alix A, Ashauer R, van den Brink P, Chapman P, Ducrot V, Forbes V, Grimm V, Schäfer D, Streissl F, Thorbek P. 2009b. Mechanistic effect models for ecological risk assessment of chemicals (MEMoRisk)-a new SETAC-Europe Advisory Group. *Environ Sci Pollut Res* 16:250-252.

Schmolke A, Thorbek P, DeAngelis DL, Grimm V. 2010. Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology & Evolution* 25:479-486.

Sousa T, Domingos T, Kooijman SALM. 2008. From empirical patterns to theory: a formal metabolic theory of life. *Phil Trans R Soc B* 363:2453-2464.

US-EPA. 2009. Guidance on the development, evaluation, and application of environmental models. U.S. Environmental Protection Agency, EPA/100/K-09/003

Van Leeuwen CJ, Vermeire TG. 2007. Risk assessment of chemicals. An introduction. Springer, Dordrecht, The Netherlands

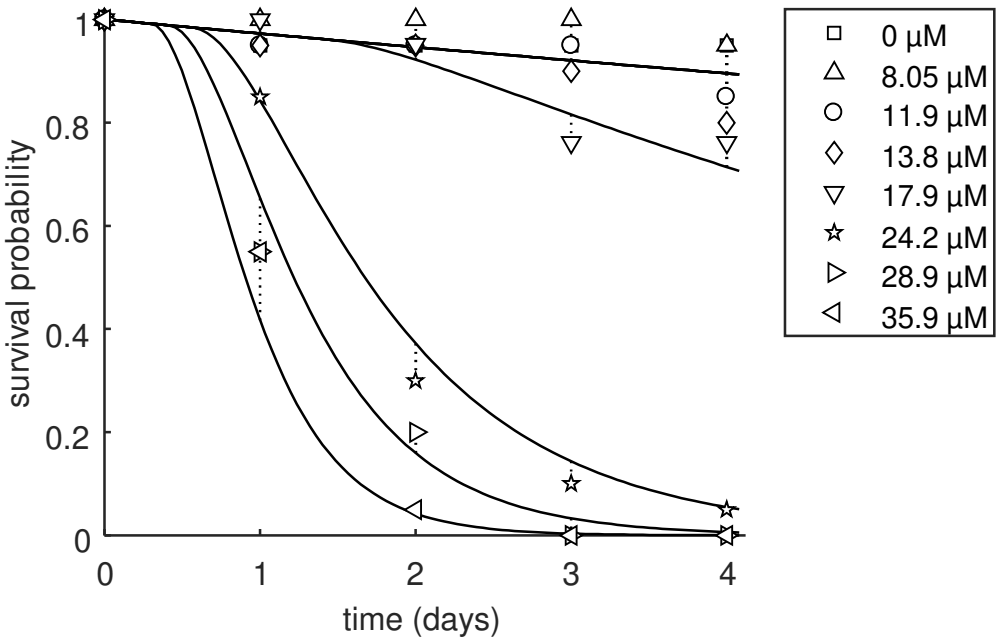

Figure 1. Fit of GUTS (reduced stochastic death model) to survival data for propiconazole in *Gammarus pulex* (Nyman et al. 2012). This fit was achieved with four parameters (95% confidence interval in parentheses): dominant rate constant 2.2 (1.6-3.3) d^{-1}, threshold for effects 17 (16-18) µM, killing rate 0.13 (0.087-0.20) $µM^{-1}$ d^{-1}, background hazard rate 0.028 (0.013-0.050) d^{-1}. The exposure scenario comprised four-day constant exposure to several concentrations (see legend).

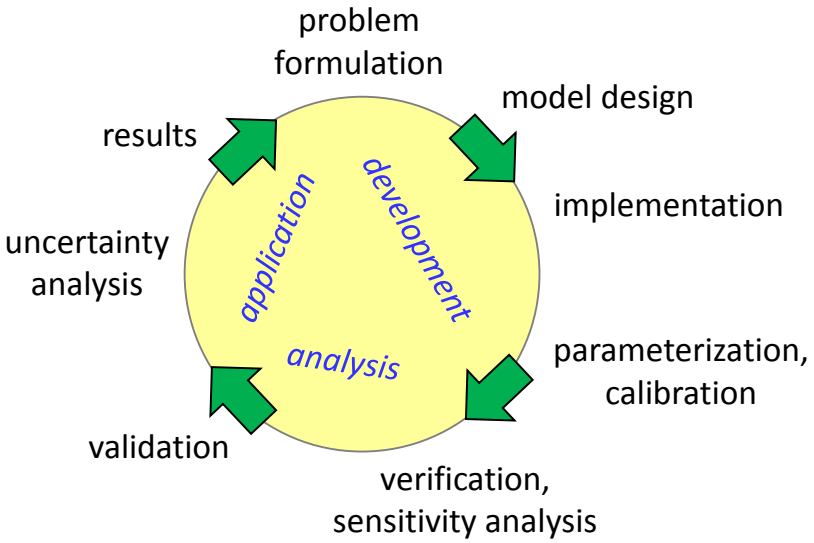

Figure 2. Simplified modelling cycle, modified from Schmolke et al. (2010).

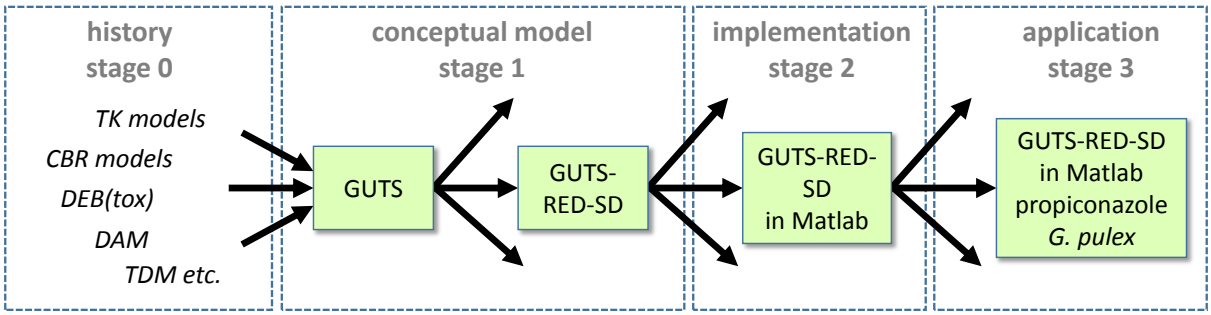

Figure 3. Different modelling stages for GUTS. The general GUTS framework has a number of special cases (such as the reduced stochastic-death model, GUTS-RED-SD), which have a range of implementations (such as in a toolbox for Matlab), which can have numerous applications to specific chemicals and species (such as to propiconazole in the amphipod *G. pulex* in Fig. 1).

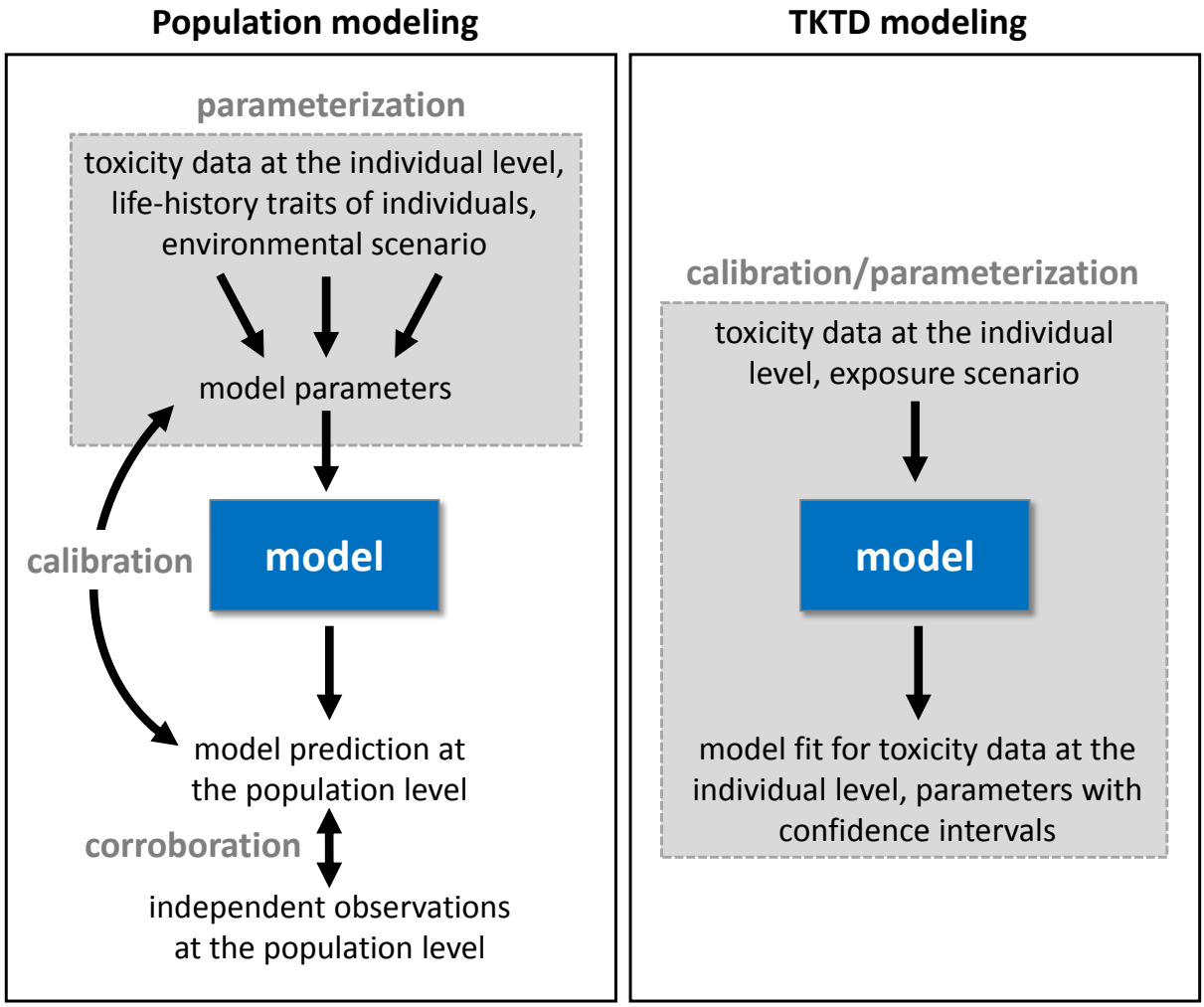

Figure 4. Difference between the typical use of population models and TKTD models. The grey box shows what 'parameterisation' involves for both types of models.

Table 1. Four different types of applications of TKTD models.

Purpose	**Explanation**	**Typical application context**
1) Analysis of data from toxicity tests	TKTD models use all of the data simultaneously: all treatments, all observation times, and in some cases all endpoints. This is more robust than fitting a dose-response curve, and can easily accommodate non-standard data.	Output from a TKTD model can be a model parameter to be used as summary statistic (such as a time-independent effect threshold), or a more standard EC*x*,*t* (exposure concentration for *x*% effect after exposure duration *t*).
2) Extrapolation to untested conditions	A TKTD model, once calibrated on toxicity data (see point 1), can be used to make predictions for other situations. As these models are based on mechanistic insights, their predictions are more useful than those based on dose-response curves.	An application with particular relevance for PPPs is extrapolation from constant to time-varying exposure (e.g., using the results from fate models). One can also think of extrapolation to different environmental conditions, such as temperature and food availability.
3) Building block in population and community models	The extrapolation opportunities (see point 2) can be used to represent individual life histories in response to dynamic exposure conditions.	TKTD models can serve as the individual-level module in models for higher levels of biological organisation, and can thus be integrated into more complex models.
4) Extrapolation between chemicals and between species	TKTD models are based on mechanisms, and explicitly separate TK from TD. Therefore, their model parameters are more likely to reveal patterns between species and between chemicals than traditional statistics such as EC*x*,*t*.	TKTD model parameters can be related to each other and to properties of the species and the chemical (QSARs). Such relationships can ultimately be used to predict model parameters for untested species and compounds.

Table 2. Example of typical items and questions for evaluation of TKTD models, categorised per modelling stage as depicted in Figure 3. Generally, each TKTD model with have a single set of concepts, several implementations, and many applications (Fig. 3). This table is meant as illustration and is not exhaustive.

Modelling stage	**Items for evaluation**	**Questions for evaluation**
0) History	Historical roots	Clearly documented and explained?
	Use of existing modules	Documented, evaluated conceptually (see stage 1), and appropriate?
1) Concepts	Model aim/domain	What types of question can be addressed?
	Underlying assumptions	Explicit, consistent and appropriate? What is their scientific support?
	Complexity	Appropriate level given model aim?
	Translation into mathematics	Clear and correct? Have verification steps been taken?
	Documentation	Complete and clear?
	Scientific status	What is extent/range of successful applications, (see stage 3)?
	Regulatory status	Already used/accepted for regulatory purposes, and/or mentioned in guidance documents?
2) Implementation	Code	Have verification steps been taken?
	Numerical methods	Appropriate and robust choice of methods?
	User friendliness	What is the extent/quality of the user interface?
	User manual	Is there a manual, is it complete and clear?
	Availability	Is implementation (and code) publicly available?
3) Application	Purpose	What specific question should be answered?
	Data	Are data and experimental design well described, and appropriate for model and application purpose?
	Numerical/statistical methods	If the implementation offers a choice of methods: is an appropriate selection made?
	Calibration	What is degree of correspondence of model to data, and is it sufficient for the purpose?
	Uncertainty analysis	Quantification and propagation of uncertainties? If so, which uncertainties are included, and have appropriate methods been used?
	Relevance	Does the model analysis address the application purpose?
	Realism of model predictions	Have model results been compared to independent data (corroboration)? What is the nature and degree of the deviations?